# Weiran Yao

Phone: +1 (412) 613-1327 • Email: weirayao@gmail.com • Website: https://weirayao.github.io/ • LinkedIn: linkedin.com/in/weiranyao/

## EMPLOYMENT

**Senior Research Scientist & Manager**, *Salesforce AI Research*, Palo Alto, CA — Jan 2023 – Current

Areas: *AI Agent, Multi-Agent System, Finetuning & Alignment, Data Pipeline, Prompt Optimization*

**Tech Lead for Agentic AI Incubation**. Drove cross-functional initiatives for AI systems for multi-agent, software engineering agent, and web agent, leading team of 4 scientists to develop high-quality synthetic data pipeline for Code LLMs in production and communicated insights for executive decision-making.

- Salesforce CodeGenie Agent [Blog][Code] — Aug 2024 – Current
- SlackAgents: Scalable Multi-Agent Workspace [Blog][Demo] — Jan 2024 – Current
- CRM WebAgent [Demo] — Aug 2023 – Oct 2023

LLM/SLM **Finetuning & Alignment**. Conducted post-training research to align long-context models to specialize in self-reflection of task executions. Contributed to Salesforce in-house xLAM-series agentic model development by aligning the model for function call in CRM production environment.

- Retroformer 7B – General Critic Model for Agentic Reflection [Paper][Code]
- xLAM 1B | 7B | 8x7B | 8x22B – Large Action Model for Function Call [Blog] [Code] [Models] [Report]

Conducted research on **Synthetic Data Pipeline** for LLM function call. This pipeline enabled a 7B model to **outperform several gpt-4 models** for function call on Berkeley Function-Calling Leaderboard.

- APIGen: Automated Pipeline for Generating Function-Calling Datasets [Blog][Paper][Data]
- AgentOhana: Unified Data and Training Pipeline for Effective Agent Learning [Report][Code]

**Prompt Engineering and Optimization**. Conducted research to automatically optimize the system prompt of LLM agent towards multi-objectives, e.g., accuracy, consistency, latency, and applied it to product.

- Einstein Copilot Meta-Prompt Optimization. Latency metrics improved by 48%.
- PRAct: Optimizing Principled Reasoning and Acting of LLM Agent [Paper][Code]

**AI Interpretability**. Conducted scalable sparse autoencoder research for extracting universal concepts across large models. Applied the approach for safe model alignment even with limited feedback.

- Editing Arbitrary Propositions in LLMs without Subject Labels [Paper][Code]

**Engineering Products**. Developed automatic root cause analysis algorithms for Salesforce Database Throttles with scalable, real-time anomaly detection. Developed Function Call and Structured Output API endpoints for Salesforce xLAM service based on vLLM inference backend.

- SRE Agent – developed dbCPU RCA agent to speed up incident response and investigation [Blog]
- OpenAI-Compatible Function Call + Structured Output API Endpoint

**Ph.D. Researcher**, *Carnegie Mellon University*, Pittsburgh, PA — Sep 2017 – Dec 2022

Areas: Fundamentals of *AI Interpretability*

My research focused on provable **AI Interpretablity** with sparse, disentangled autoencoders to identify concepts, and cause and effect from videos and non-stationary time series. Some selected work below.

- Temporally Disentangled Representation Learning [Paper][Code]
- Learning Temporally Causal Latent Processes from General Temporal Data [Paper][Code]
- Prompt Learning with Optimal Transport for Vision-Language Models [Paper][Code]

## OPEN-SOURCE SOFTWARE

**AgentLite:** Lightweight Library for Building LLM Multi-Agent System (538 Stars)

**CausalAI:** Scalable framework for Causal Analysis of Time Series and Tabular Data (265 Stars)

**Merlion:** A Machine Learning Framework for Time Series Intelligence (3.4k Stars)

## EDUCATION

**Carnegie Mellon University**, School of Computer Science, Pittsburgh, PA

- Ph.D. in Advanced Infrastructure Systems — Aug 2017 – Aug 2023
- M.S. in Machine Learning — Aug 2019 – May 2021

## TECH STACK

**Programming Language**: Python, JavaScript, HTML/CSS, Bash, SQL

**Tools and Frameworks**: PyTorch, Docker, Kubernetes, Streamlit, FastAPI, Git, LaTeX

## PUBLICATIONS

### CONFERENCE AND JOURNAL PUBLICATIONS

[Google Scholar is Here]

[27] SpecTool: A Benchmark for Characterizing Errors in Tool-Use LLMs

[26] Language Models are Hidden Reasoners: Unlocking Latent Reasoning Capabilities via Self-Rewarding

[25] **PRACT: Optimizing Principled Reasoning and Acting of LLM Agent**

*Empirical Methods in Natural Language Processing, 2024.*

[24] **xLAM: A Family of Large Action Models to Empower AI Agent Systems**

[23] **Diversity Empowers Intelligence: Integrating Expertise of Software Engineering Agents.**
*ACM International Conference on Information and Knowledge Management (CIKM), 2024.*

[22] **APIGen: Automated Pipeline for Generating Verifiable and Diverse Function-Calling Datasets.**
*Advances in Neural Information Processing Systems (NeurIPS), 2024.*

[21] AgentOhana: Design Unified Data and Training Pipeline for Effective Agent Learning.

[20] AgentLite: A Lightweight Library for Building and Advancing Task-Oriented LLM Agent System.

[19] CaRiNG: Learning Temporal Causal Representation under Non-Invertible Generation Process.
*International Conference on Machine Learning (ICML) 2024.*

[18] Causal Layering via Conditional Entropy.
*Causal Learning and Reasoning (CLeaR) 2024.*

[17] Editing Arbitrary Propositions in LLMs without Subject Labels.

[16] DRDT: Dynamic Reflection with Divergent Thinking for LLM-based Sequential Recommendation.

[15] Temporally Disentangled Representation Learning under Unknown Nonstationarity.
*Advances in Neural Information Processing Systems (NeurIPS), 2023.*

[14] **Retroformer: Retrospective Large Language Agents with Policy Gradient Optimization**.
*International Conference on Learning Representations (ICLR) 2024.* ***(Spotlight Presentation)***.

[13] BoLAA: Benchmarking and Orchestrating LLM-Augmented Autonomous Agents.
*International Conference on Learning Representations (ICLR) 2024.*

[12] Rex: Rapid Exploration and Exploitation for AI Agents.
*International Conference on Learning Representations (ICLR) 2024.*

[11] On the Unlikelihood of D-Separation.
*The International Conference on Probabilistic Graphical Models (PGM) 2024.*

[10] Salesforce CausalAI Library: A Fast and Scalable Framework for Causal Analysis of Time Series and Tabular Data.

[9] Non-Parametric State-Space Models: Identifiability, Estimation and Forecasting.
*International Conference on Learning Representations (ICLR) 2023.*

[8] Temporally Disentangled Representation Learning.
*Advances in Neural Information Processing Systems (NeurIPS), 2022.*

[7] **Prompt Learning with Optimal Transport for Vision-Language Models.**
*International Conference on Learning Representations (ICLR) 2023.* ***(Spotlight Presentation)***.

[6] **Distribution-aware Goal Prediction and Model-based Planning for Safe Autonomous Driving.**
*International Conference on Machine Learning (ICML) 2022. Workshop on Safe Learning for Autonomous Driving* ***(Best Paper Award)***.

[5] Partial Disentanglement for Domain Adaptation.
*International Conference on Machine Learning (ICML) 2022.*

[4] Learning Temporally Causal Latent Processes from General Temporal Data.
*International Conference on Learning Representations (ICLR) 2022.*

[3] Data Driven Safety Risk Prediction of Lithium Ion Battery.
*Advanced Energy Materials 2021.*

[2] From Twitter to traffic predictor: Next-day morning traffic prediction using social media data.
*Transportation Research Part C: Emerging Technologies 2021.*

[1] Learning a Distributed Control Scheme for Demand Flexibility in Thermostatically Controlled Loads.
*IEEE SmartGridCommm. 2020.*

**PATENTS**

[11] Systems And Methods For Function-Calling Agent Models, US Patent, 636,605,12

[10] Systems And Methods For Building a Code Generation Agent, US Patent, 636,815,24

[9]Systems And Methods For Building Task-Oriented Hierarchical Agent Architectures, US Patent, 187,389,84

[8] Systems And Methods For Controllable Artificial Intelligent Agents, US Patent, 188,170,64

[7] Systems And Methods For Language Agent Optimization, US Patent 18,498,257.

[6] Systems And Methods For Orchestrating LLM-Augmented Autonomous Agents, US Patent 18,494,393.

[5] Systems And Methods For Building AI Agents For Language Models, US Patent 63,555,382.

[4] Systems And Methods For A Unified Training Framework Of Large Language Models, US Patent 18,658,899.

[3] Systems And Methods For Editing A Large Language Model, US Patent 18,428,530.

[2] Systems And Methods For A Unified Training Framework Of Large Language Models, US Patent 18,658,899.

[1] Distributed Control for Demand Flexibility in Thermostatically Controlled Loads, US Patent 12,027,858.

### PRESS COVERAGE

[9] **VentureBeat**. "Is AI the future of sales? Salesforce's new models could change the game."

[8] **TimesOfAI**. 'Salesforce DEI: How Diversity Is Driving AI Innovation in Software Engineering."

[7] **MarkTechPost**. "Salesforce AI Research Proposes DEI: AI Software Engineering Agents Org, Achieving a 34.3% Resolve Rate on SWE-Bench Lite, Crushing Closed-Source Systems."

[6] **VentureBeat**. "Salesforce proves less is more: xLAM-1B 'Tiny Giant' beats bigger AI Models."

[5] **The Stack**. "On-device agentic AI is here!"

[4] **MarkTechPost**. "Salesforce Research Introduces AgentOhana: A Comprehensive Agent Data Collection and Training Pipeline for Large Language Model."

[3] **MarkTechPost**. "AgentLite by Salesforce AI Research: Transforming LLM Agent Development with an Open-Source, Lightweight, Task-Oriented Library for Enhanced Innovation."

[2] **MarkTechPost**. "Salesforce AI Researchers Introduce the Evolution of LLM-Augmented Autonomous Agents and the Innovative BOLAA Strategy."

[1] **MarkTechPost**. "Meet Retroformer: An Elegant AI Framework for Iteratively Improving Large Language Agents by Learning a Plug-in Retrospective Model."

### INDUSTRY TALKS

[3] **Large Actions Models in a Multi-Agent World, Breakout Session at *Dreamforce 2024*, Sep 2024, San Francisco.**

[2] PRAct: Optimizing Principled Reasoning and Acting of LLM Agent, invited talk at *Databricks Data + AI Summit*, Jun 2024, San Francisco.

[1] Retroformer: Retrospective Large Language Agents with Policy Gradient Optimization, invited talk at *Moveworks*, Sep 2023, Mountain View.

### BLOGS

[2] Meet Merlion: An End-to-End Easy-to-Use Machine Learning Library for Time Series Applications. Salesforce AI Research.

[1] CausalAI: Answering Causality Questions Using Observational Data. Salesforce AI Research.

## MENTORING EXPERIENCE

**Summer Intern @ Salesforce AI Research**

- Kexun Zhang, Ph.D. student at Carnegie Mellon University, Language Technology Institute.

**Ph.D. Student @ Carnegie Mellon University**

- Lingjing Kong, Ph.D. student at Carnegie Mellon University Machine Learning Department.
- Xiangchen Song, Ph.D. student at Carnegie Mellon University Machine Learning Department.
- Zemian Ke, Ph.D. student at Carnegie Mellon University Mobility Data Analytics Center.